DEEPSEA: A Multilingual Evaluation Benchmark for Southeast Asia

Jiuzhou Han* **Abhishek Bijay Mishra** **Trevor Cohn** **Shubham Mittal**
Google Research
jiuzhou.han@monash.edu, {abhishekm, tcohn, shmittal}@google.com

Abstract

While language models have demonstrated remarkable capabilities, their performance remains disproportionately skewed towards high-resource languages. This disparity is particularly acute for the 700 million inhabitants of Southeast Asia, where vast linguistic diversity remains under-represented in existing benchmarks, often suffering from limited language coverage or narrow task scopes. To address this, we introduce DEEPSEA, a comprehensive benchmark constructed from established datasets through professional human translation into 12 Southeast Asian languages, spanning both high-resource (e.g., Indonesian, Thai, Vietnamese) and low-resource (e.g., Burmese, Khmer, Lao) tiers across five distinct tasks: mathematical reasoning, scientific reasoning, question answering, abstractive summarization, and machine translation. Extensive evaluation of frontier proprietary and open-weights models reveals that while frontier models like Gemini-3.7-Flash maintain robustness across resource tiers, smaller models face severe performance degradation on low-resource languages. Crucially, models perform better in reasoning and summarization when pivoted through English rather than target languages, exposing a critical dependency on English representations, especially for low-resource languages. Beyond these observations, we provide a cross-lingual error analysis and a cross-script tokenizer fertility analysis. We hope these findings shed light on the development of next-generation inclusive multilingual models.[1]

1 Introduction

The advent of Large Language Models (LLMs) has revolutionized natural language processing, demonstrating remarkable capabilities in reasoning, generation, and understanding (OpenAI, 2023; Gemini Team, 2025; DeepSeek-AI, 2025). However, despite these advancements, the benefits of LLMs remain disproportionately skewed towards high-resource languages, particularly English and European languages. This disparity creates a barrier for the 700 million inhabitants of Southeast Asia (SEA), a region characterized by high linguistic diversity and a complex tapestry of high, medium, and extremely low-resource languages (Susanto et al., 2025; Lovenia et al., 2024).

While recent initiatives have evaluated multilingual LLMs across the SEA region, existing benchmarks often suffer from limited language coverage or narrow task scopes. Leading benchmarks like SeaEval (Wang et al., 2024), SeaExam, and SeaBench (Liu et al., 2025) introduce valuable frameworks for cultural reasoning and local examinations, yet cluster disproportionately around major national languages (primarily Indonesian, Vietnamese, and Thai). SEA-HELM (Susanto et al., 2025) introduces a holistic five-pillar framework but evaluates only five languages, leaving vast portions of the region unrepresented. Furthermore, while SEACrowd (Lovenia et al., 2024) consolidates resources for nearly 1,000 languages, it serves primarily as a data hub rather than a targeted evaluation suite across the resource spectrum.

A critical gap persists in the lack of a unified evaluation benchmark covering both major national languages and underserved low-resource regional languages. To bridge this gap, we introduce **DEEPSEA**, the first comprehensive, human-translated benchmark designed to rigorously evaluate LLMs across 12 Southeast Asian languages. DEEPSEA assesses capabilities in understanding, multi-step reasoning, and generation across five distinct tasks: mathematical reasoning (MGSM), scientific reasoning (IndoMMLU), question answering (XQuAD), abstractive summarization (CrossSum), and machine translation (WMT24++).

*Work done during an internship at Google.

[1] DEEPSEA is open source and available at https://github.com/google-research/deep-sea.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 1–17
October 24-29, 2026 ©2026 Association for Computational Linguistics

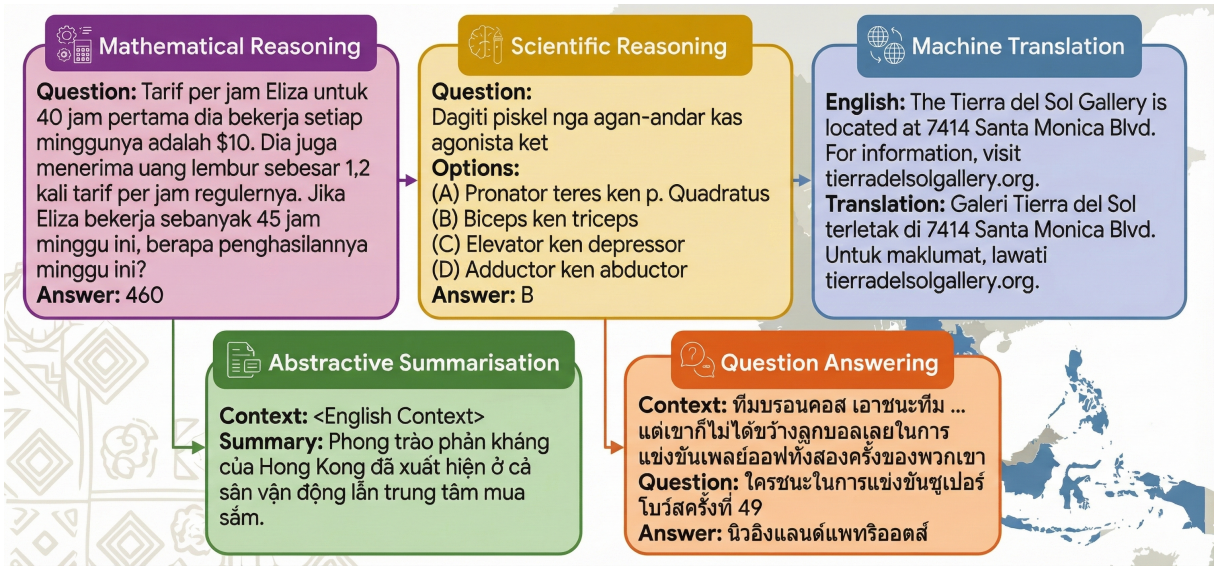

Figure 1: Demonstration of various tasks covered in DEEPSEA, illustrating specific examples of inputs and outputs.

Our contributions are as follows:

- We introduce DEEPSEA, a human-translated benchmark spanning 12 SEA languages across high- and low-resource tiers, evaluated on five understanding and generation tasks.
- We provide a systematic evaluation of the LLM landscape, ranging from frontier proprietary systems to open-weights models.
- We conduct a granular performance analysis across languages, isolating key bottlenecks and providing actionable insights for future AI development for the SEA region.

2 Related Work

High-Coverage Parallel Benchmarks Multilingual evaluation has traditionally relied on N-way parallel corpora for cross-lingual comparison (Conneau et al., 2018). The translation evaluation dataset, FLORES-200 (Costa-jussà et al., 2022) and its reading comprehension derivative Belebele (Bandarkar et al., 2024) are the only benchmarks covering all 12 SEA languages examined in this study to the best of our knowledge. Topic classification datasets SIB-200 (Adelani et al., 2024) and Taxi1500 (Ma et al., 2025) also cover all 12 languages, but restrict labels to English. Other high-coverage parallel datasets include NTREX-128 (Federmann et al., 2022) (translation) and MASSIVE (FitzGerald et al., 2023) (slot filling), both covering 8 of the 12 languages. However, the task diversity in these datasets is limited.

Global Knowledge and Reasoning Benchmarks The above parallel datasets are translated from an English base, but recent efforts have shifted to more globally representative benchmarks. Global-MMLU (Singh et al., 2025) translates only the subset of MMLU that is culture-agnostic. In contrast, non-parallel benchmarks like MultiLoKo (Hupkes and Bogoychev, 2025), TyDiQA (Clark et al., 2020), and M3exam (Zhang et al., 2023) are built using locally sourced native questions across diverse languages. Similarly, INCLUDE (Romanou et al., 2025) and Global-PIQA (Chang et al., 2025) leverage large-scale crowdsourcing to collect high-coverage, culturally grounded items. While all these global datasets cover a few SEA languages, most of the 12 languages examined in this work are missing from each, with Sundanese, Ilocano, and Cebuano being present in none, despite accounting for 60 million speakers in aggregate.

SEA Regional Benchmarks Recent evaluation frameworks for Southeast Asia span both regional and dialect-specific scopes. At the regional level, SeaEval (Wang et al., 2024), SeaExam, and SeaBench (Liu et al., 2025) emphasize cultural reasoning and local examinations, while SEA-HELM (Susanto et al., 2025) introduces a holistic five-pillar framework, though limited to five languages. Significant attention has also been directed towards Indonesia's linguistic diversity; IndoMMLU (Koto et al., 2023) targets academic proficiency, while LoraxBench (Aji and Cohn, 2025) and NusaX (Winata et al., 2023) provide extensive coverage of local

Language	Code	Family	Script	Typology
Vietnamese	vi	Austroasiatic	Latin	Analytic
Filipino	fil	Austronesian	Latin	Agglutinative
Indonesian	id	Austronesian	Latin	Agglutinative
Thai	th	Tai-Kadai	Thai	Analytic
Malay	ms	Austronesian	Latin	Agglutinative
Cebuano	ceb	Austronesian	Latin	Agglutinative
Javanese	jv	Austronesian	Latin	Agglutinative
Sundanese	su	Austronesian	Latin	Agglutinative
Burmese	my	Sino-Tibetan	Myanmar	Agglutinative
Ilocano	ilo	Austronesian	Latin	Agglutinative
Khmer	km	Austroasiatic	Khmer	Analytic
Lao	lo	Tai-Kadai	Lao	Analytic

Table 1: Linguistic characteristics of the covered languages. The first four languages are classified as high-resource, while the remaining eight are low-resource.

languages through translated and parallel tasks. Beyond Indonesia, specialized benchmarks target other major national languages, such as BATAYAN (Montalan et al., 2025) for Filipino, ThaiCLI (Kim et al., 2025) for Thai cultural nuances, and ViQuAD (Van Nguyen et al., 2020) for Vietnamese.

3 DEEPSEA

3.1 Languages of Focus

The benchmark targets 12 Southeast Asian languages, stratified into two resource categories, high and low. Group high comprises Thai, Indonesian, Vietnamese, and Filipino. Group low includes Cebuano, Ilocano, Javanese, Khmer, Lao, Malay, Sundanese, and Burmese. Table 1 demonstrates the linguistic characteristics of these languages.

3.2 Task Coverage

To ensure a comprehensive evaluation, DEEPSEA encompasses five distinct tasks derived from established benchmarks. Where target-language coverage was unavailable, we commissioned professional native translators to translate the original instances into the target Southeast Asian languages. Figure 1 illustrates examples of the five tasks.

Mathematical Reasoning To evaluate arithmetic reasoning ability, we adopt data from the Multilingual Grade School Math (MGSM) (Shi et al., 2023). MGSM contains grade-school level math problems that require multi-step reasoning to solve.

Scientific Reasoning Our scientific reasoning dataset is derived from IndoMMLU (Koto et al., 2023), a massive multitask language understanding benchmark in Indonesian. We specifically leveraged a subset of questions focusing on STEM domains: Biology, Chemistry, and Physics.

Question Answering For this task, we use XQuAD (Artetxe et al., 2020), consisting of reading comprehension passages accompanied by questions that require extractive answers.

Abstractive Summarization We assess summarization capabilities using CrossSum (Bhattacharjee et al., 2023), a large-scale cross-lingual summarization corpus. This task challenges models to generate a concise summary in a target language based on the provided source document in English.

Machine Translation The machine translation dataset is constructed based on data from WMT shared tasks (Kocmi et al., 2024; Deutsch et al., 2025). We focus specifically on the translation direction from English to the target languages to assess the models' generative capabilities in low-to-high resource settings.

3.3 Data Creation and Statistics

We aggregated existing publicly available data and extended the benchmark by generating new translations for unsupported languages. We commissioned professional native translators to map the data from the respective source languages to the target SEA languages. For MGSM, XQuAD, WMT, and CrossSum, translators worked directly from the English source text. Since IndoMMLU is originally an Indonesian benchmark, translations for this task were conducted from Indonesian into the target low-resource languages (details in Appendix A).

To preserve the meaning and evaluation format of the original instances, translators were provided with explicit translation guidelines such as translating only the designated text field, while using the original article or context, when available, solely as a reference for disambiguation. They were asked to provide a single natural translation rather than multiple alternatives and to avoid adding clarifications, redundant information, or explanatory content not present in the source. Numerical expressions were required to retain the same form as the original input (e.g., words versus numerals), while punctuation, quotation marks, capitalization, and the overall formatting of the source text were preserved as closely as possible. To ensure high translation quality and ethical labour standards, the entire DEEPSEA dataset was translated through professional translation vendors, with all contributors receiving fair, competitive compensation well above local minimum wage guidelines (e.g., $0.16 per word in MGSM).

Task	Eval	# high	# low
Mathematical Reasoning	MGSM	1,000	2,000
Scientific Reasoning	IndoMMLU	2,940	5,880
Question Answering	XQuAD	4,760	9,520
Abstractive Summarization	CrossSum	1,558	3,569
Machine Translation	WMT24++	3,988	7,976

Table 2: Number of samples in DEEPSEA eval sets, grouped by 4 high and 8 low-resource languages.

Quality check of the DEEPSEA benchmark confirms the high fidelity and accuracy of the translated datasets (refer to Appendix B for more details). Table 2 summarizes the statistics of various datasets. Except for CrossSum, all tasks maintain a strict one-to-one parallel alignment with English; in CrossSum, splits for Thai (th), Indonesian (id), Vietnamese (vi), and Burmese (my) lack direct counterparts because matching source articles were unavailable in the original corpus.

4 Experiments

4.1 Models

We evaluate a diverse spectrum of models across three categories. Firstly, the frontier proprietary systems comprising Gemini-3.7-Flash (Google DeepMind, 2026), Gemini-3-Pro (Google DeepMind, 2025b), Gemini-3-Flash (Google DeepMind, 2025a), GPT-5.6-Luna (OpenAI, 2026), and Claude-Sonnet-5 (Anthropic, 2026). Secondly, the leading open-weights models, including Gemma-3 (Gemma Team, 2025), Qwen2.5 (Yang et al., 2024), and Llama-3.1 (Llama Team, 2024). Lastly, we also evaluate the regionally adapted models specifically optimised for Southeast Asian languages, namely SEA-LION (Ng et al., 2025) and Sailor2 (Dou et al., 2025). Further, additional model results are provided in Appendix Table 6.

4.2 Inference Setup

All models are evaluated in a zero-shot setting with greedy decoding (temperature $= 0$) to ensure deterministic results. For reasoning-intensive tasks (i.e., mathematical and scientific reasoning), we explicitly instruct models to generate step-by-step reasoning. To maintain linguistic consistency, models are directed to reason in the same language in which the problem is presented. Complete prompt templates for each task are provided in Appendix D.

4.3 Evaluation Metrics

We employ task-specific metrics across the DEEPSEA benchmark. For reasoning-intensive tasks (mathematical and scientific reasoning), we report standard accuracy, and for question answering, we measure token-level F1. For abstractive summarization, we report chrF (Popovic, 2015), following prior multilingual benchmarks (Montalan et al., 2025). Finally, machine translation performance is assessed using MetricX-24 (Juraska et al., 2024). Table 3 summarizes the results across all evaluated models and tasks.

4.4 Results

4.4.1 Results Across Models

Frontier Proprietary Models Gemini-3.7-Flash consistently achieves state-of-the-art performance across all evaluated tasks and demonstrates particular robustness in maintaining high accuracy across language tiers, exhibiting minimal degradation from English to high- and low-resource languages. For instance, in mathematical reasoning (MGSM), Gemini-3.7-Flash maintains a score of 96.4 in the low-resource category, significantly outperforming the best open-weights model, SEA-LION-v4-27B, which scores 83.8. Claude-Sonnet-5 is also highly competitive, particularly in mathematical reasoning and question answering, while GPT-5.6-Luna also performs strongly on the reasoning tasks and ties Gemini-3.7-Flash for the best machine translation performance.

Open-Weights Models Among general-purpose open-weights models, we observe a clear scaling trajectory where increased model capacity substantially strengthens cross-lingual transfer. Gemma-3-27B consistently outperforms Gemma-3-12B across tasks, achieving 71.5 accuracy on low-resource scientific reasoning compared to 63.6 for the 12B variant. In contrast, the smaller generalist models (Qwen2.5-7B-Inst and Llama-3.1-8B-Inst) suffer from a severe performance collapse on low-resource languages despite respectable English accuracy. In mathematical reasoning, Qwen2.5-7B-Inst drops from 93.2 in English to 78.3 (high) and 36.2 (low)—a degradation of over 60%—whereas the frontier proprietary models or open-weights regional models maintain much flatter cross-lingual curves. This highlights that the small language models lag behind in their multilingual capabilities, unless specialised for the target languages.

Model	Mathematical Reasoning			Scientific Reasoning			Question Answering			Abstractive Summarization			Machine Translation	
	Accuracy ↑			Accuracy ↑			F1 ↑			ChrF ↑			MetricX-24 ↓	
Language	en	high	low	en	high	low	en	high	low	en	high	low	high	low
Gemini-3.7-Flash	**99.2**	**96.5**	**96.4**	**86.5**	**86.3**	**86.1**	**87.9**	**85.2**	**81.7**	**37.4**	**36.4**	**36.2**	**2.4**	**3.5**
Gemini-3-Pro	98.0	95.6	95.4	85.0	85.3	84.3	85.6	81.2	77.6	37.0	35.9	35.4	2.5	3.7
Gemini-3-Flash	98.8	95.7	95.7	85.7	85.2	84.5	87.2	83.6	79.5	36.9	36.2	**36.2**	2.5	3.7
GPT-5.6-Luna	96.4	93.3	92.8	85.3	84.5	83.3	73.8	68.1	65.8	33.7	33.8	33.2	**2.4**	**3.5**
Claude-Sonnet-5	98.8	96.4	95.4	84.9	84.6	82.6	87.1	84.3	80.3	35.7	35.8	34.0	2.6	3.8
Gemma-3-27B	96.0	91.7	83.6	79.2	76.2	71.5	85.5	79.9	73.9	33.1	33.5	29.7	2.9	6.1
Gemma-3-12B	92.8	89.7	77.2	71.0	68.5	63.6	84.7	77.3	71.5	33.4	32.5	27.5	3.1	8.6
Qwen2.5-7B-Inst	93.2	78.3	36.2	68.2	61.3	49.2	71.9	48.5	37.9	32.9	28.4	22.2	8.2	17.6
Llama-3.1-8B-Inst	87.6	67.3	38.8	52.0	42.3	24.5	71.1	62.9	37.2	34.8	30.3	21.5	5.9	14.8
SEA-LION-v4-27B	96.0	92.7	83.8	78.1	76.2	71.5	85.6	80.1	74.4	33.0	33.0	28.8	3.0	6.0
Sailor-2-20B	92.8	87.7	77.6	51.3	37.9	34.3	74.7	57.4	46.1	31.4	30.4	27.2	4.6	12.6

Table 3: Performance of selected models across different tasks, with languages split among English, high- and low-resource groups as defined in Table 1. **Bold** values indicate the best overall performance, while underlined values indicate the best performance among open-weights models.

Regional Models The performance of models specialised for SEA region, such as SEA-LION-v4-27B, shows impressive competitiveness against the generalist models of similar size, particularly in high-resource languages. It effectively matches or surpasses Gemma-3-27B in XQuAD and WMT24++, suggesting that regionally pre-trained models remain a vital strategy for optimising performance in specific linguistic geographies.

4.4.2 Results Across Tasks

Reasoning Tasks (MGSM and IndoMMLU) Mathematical and scientific reasoning appear to be the most discriminative tasks for evaluating cross-lingual capability. While English performance is relatively high across most models, the gap widens considerably for low-resource languages. Gemini-3.7-Flash demonstrates particularly strong robustness, decreasing only from 99.2 to 96.4 on Mathematical Reasoning and from 86.5 to 86.1 on Scientific Reasoning between English and low-resource languages. Among general-purpose open-weights models, Gemma-3-27B achieves the strongest low-resource performance, with scores of 83.6 and 71.5 on the two reasoning tasks, respectively. This suggests that although reasoning ability may transfer across languages, linguistic accessibility remains a major bottleneck for weaker models.

Question Answering (XQuAD) Performance on XQuAD exhibits a sharper cross-lingual drop between English and SEA languages than the reasoning tasks. Gemini-3.7-Flash maintains the strongest performance, decreasing from 87.9 in English to 81.7 on low-resource languages. Notably, GPT-5.6-Luna performs substantially lower on question answering relative to other tasks, achieving F1 scores of 73.8, 68.1, and 65.8 across English, high-resource, and low-resource tiers, respectively. Our error analysis reveals that this drop is driven by poor instruction-following: while the prompt strictly requires answer substring extraction from the context, the model often outputs paraphrased responses and gets penalised by token-level F1.

Abstractive Summarization (CrossSum) The CrossSum task reveals relatively stable performance across language tiers for the strongest Gemini models. Gemini-3.7-Flash achieves chrF scores of 37.4, 36.4, and 36.2 for English, high-resource, and low-resource languages, respectively, while Gemini-3-Flash also achieves 36.2 chrF on low-resource languages. However, the low-resource tier remains a significant bottleneck for smaller generalist models, with Qwen2.5-7B-Inst and Llama-3.1-8B-Inst dropping to 22.2 and 21.5, respectively. Among general-purpose open-weights models, Gemma-3-27B achieves the strongest low-resource result of 29.7 chrF, slightly outperforming the regional SEA-LION-v4-27B at 28.8 chrF.

Machine Translation (WMT24++) Gemini-3.7-Flash and GPT-5.6-Luna show the best performance, achieving MetricX-24 scores of 2.4 and 3.5 for high- and low-resource languages, respectively. Among open-weights models, Gemma-3-27B achieves the strongest high-resource score of 2.9 and remains competitive with SEA-LION-v4-27B on low-resource languages with a score of 6.1. Smaller models struggle significantly more, with

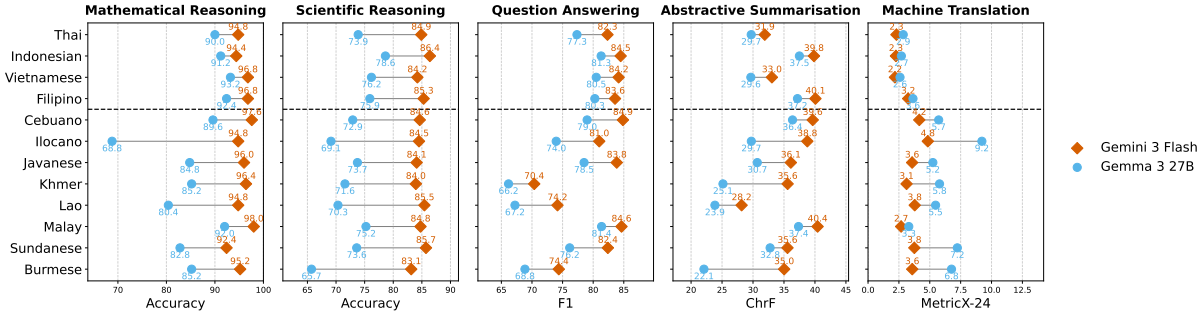

Figure 2: Frontier models maintain robustness across resource tiers, while open-weights models show severe degradation on low-resource languages. Gemini-3-Flash and Gemma-3-27B are denoted by orange diamonds and blue circles, respectively. The dotted line separates high- and low-resource languages.

Qwen2.5-7B-Inst and Llama-3.1-8B-Inst reaching scores of 17.6 and 14.8, resp., in low-resource languages. This reveals that translation quality for under-represented languages strongly differentiates the frontier and open-weights models.

4.4.3 Results across Languages

Figure 2 shows the scores across 12 SEA languages, using Gemini-3-Flash and Gemma-3-27B as the representative frontier and open-weights systems.

Near-Parity in High-Resource Languages Both the models perform consistently well across the high-resource languages, maintaining high scores on all five tasks. Across all 12 languages, Gemini-3-Flash nearly achieves cross-lingual parity, maintaining $> 92\%$ in math and $> 83\%$ in science. In contrast, Gemma-3-27B fluctuates widely, suffering steep drops on low-resource languages, showing that cross-lingual parity remains a key strength of the frontier models.

Disparity in Low-Resource Languages Low-resource languages show substantial performance degradation, especially in non-Latin scripts: Burmese, Khmer, and Lao. These three languages yield the lowest scores in question answering (< 75 F1) and summarization (e.g., 22.1 chrF for Burmese in Gemma-3-27B). Regional Latin-script languages like Ilocano also suffer sharp drops (68.8% in math and 9.2 in translation), revealing that data scarcity and script differences remain major hurdles for open-weights models.

5 Analysis

5.1 Language of Reasoning

To investigate the effect of language in reasoning-intensive tasks, we evaluate models on MGSM and IndoMMLU using two prompting methodologies. In the *English CoT* method, models receive an English-language prompt that explicitly instructed them to provide their step-by-step reasoning in English. Whereas in the *native CoT* method, the entire prompt is translated into the language of each test instance, and models are explicitly instructed to provide their reasoning in that language (e.g., in Thai for Thai instances). These prompting methods differ from the default task prompts used in Section 4.4, which have the general instruction to reason in the same language as the problem. For brevity, Appendix D presents only the English versions of the prompt templates for the mathematical and scientific reasoning tasks as representative examples; the corresponding native CoT prompts were fully translated into each target language. The results are provided in Figure 3.

Improvements using English CoT Across both the benchmarks, models using the English CoT method consistently achieve better or equal accuracy than the native CoT prompt. This shows that models often perform better when reasoning instructions are provided in English and they are explicitly asked to reason in English. While English CoT prompting may help mitigate linguistic bottlenecks associated with target-language prompting, it does reveal that models are less capable to perform when asked in languages other than English.

Performance Gradient in Native CoT Performance gap between native and English CoT seems to exacerbate from high- to low-resource languages. In IndoMMLU, Gemma-3-12B and SEA-LION-v4 get performance jump when using English CoT. This further supports earlier observation that the models are less likely to perform when natively thinking in low-resource languages.

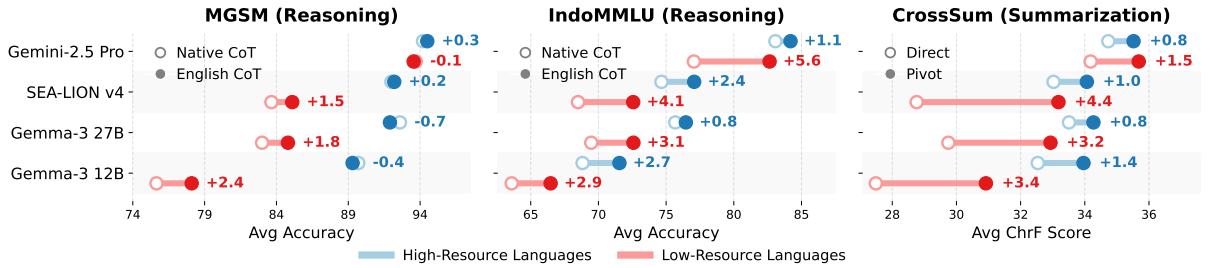

Figure 3: Reasoning in English yields consistently higher accuracy than native-language reasoning, especially for low-resource languages. The two left panels compare English and native CoT on reasoning tasks, while the right panel compares direct and English-pivot summarisation.

Task Sensitivity The benefit of English CoT appears more distinct in scientific reasoning (IndoMMLU) than in mathematical reasoning (MGSM). In MGSM, the performance bars for native and English CoT are often closer in height, possibly because MGSM is an easier task than IndoMMLU. This may also suggest that mathematical reasoning is less sensitive to the choice of CoT prompting language, whereas scientific reasoning appears more sensitive to the difference between English and native CoT prompting.

Frontier Model Robustness Gemini-2.5-Pro exhibits the highest degree of stability between reasoning strategies. In MGSM, its Native CoT performance is nearly identical to its English CoT performance across both high and low tiers. This implies that frontier models possess strong multilingual reasoning capabilities, reducing the necessity of an English pivot compared to open-weights models.

5.2 Summarisation Strategy

To evaluate the efficacy of cross-lingual information abstraction, we conduct experiments on the CrossSum dataset using two prompting strategies. The first, "Direct", instructs the model to generate the summary in the target language directly from the English source. The second, "Pivot-based", explicitly decomposes the task: first summarising the English source into an English summary and then translating that summary into the target language. The prompt templates are given in Appendix D.

As shown in the results in Figure 3, the benefits of the Pivot-based decomposition are clearly evident for low-resource languages. The performance delta is consistently wider when compared to high-resource languages. For instance, SEA-LION-v4-27B shows a large +4.4 chrF boost for low-resource languages whereas high-resource languages get +1.0 lift. This indicates that while models may possess sufficient internal knowledge to summarise directly into high-resource languages, their abstractive capabilities in low-resource settings remain brittle. The Pivot strategy effectively converts a complex abstractive task into a simpler translation task, which current models handle with greater robustness.

We attribute these gains to the effective separation of information compression (summarisation) from linguistic production (translation). By leveraging English as a high-fidelity intermediate representation, the Pivot approach secures the semantic core before addressing the perplexity challenges of the target language. This strategy mitigates the task interference typical of the Direct setting, where the dual burden of summarising and translating simultaneously often leads to content loss or hallucinations in low-resource languages.

5.3 Error Analysis

Beyond metrics, we conducted a granular analysis of failure modes in low-resource languages. We focus on two critical phenomena: Cross-Lingual Lexical Interference in reasoning tasks and Contextual Hallucinations in extractive QA tasks.

Cross-Lingual Interference While investigating incorrect responses in the mathematical reasoning task, we identified a specific class of errors driven by cross-lingual lexical interference. This occurs when the model correctly identifies a lexical token but resolves its semantics based on a related, often more dominant, language within the same family.

An interesting example is observed in Ilocano. In the dataset, the term "lawas" is used to denote "week". However, in Cebuano and other Visayan languages, the same lexeme "lawas" signifies "body". In several failed Ilocano reasoning chains, we observed models generating contexts related to biological or physical health instead of

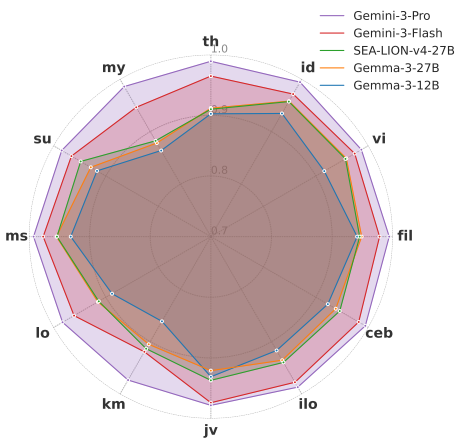

Figure 4: Radar chart illustrating the answer presence rate across 12 SEA languages in XQuAD dataset.

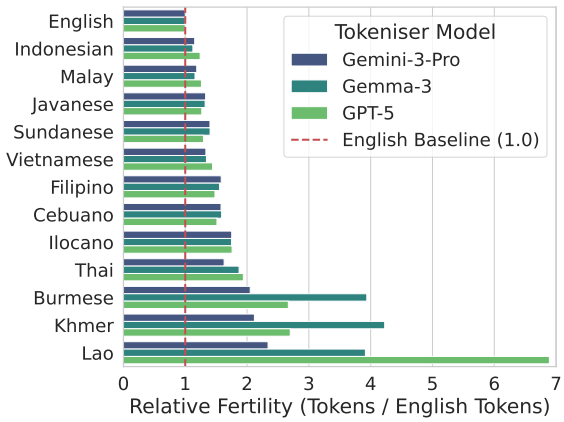

Figure 5: Relative token fertility across SEA languages when compared to English. Values represent the ratio of tokenized tokens in the given language to English tokens (baseline 1.0), calculated using three tokenizers. Higher values indicate lower tokenization efficiency.

temporal ones, suggesting the model conflated the Ilocano prompt with Cebuano semantics.

Although this specific error type is not the dominant failure mode, its presence in low-resource settings highlights a critical ambiguity issue in multilingual models. We posit that this interference risk may also exist for other linguistically proximate pairs in the region, such as Thai and Lao (Tai-Kadai family) or Indonesian and Malay (Austronesian family), where shared vocabulary often possesses divergent definitions. While likely less frequent in high-resource pairs due to stronger language identification capabilities, this semantic overlap presents a latent challenge for zero-shot inference in under-represented languages like Lao or Ilocano, where the model's internal representation may be biased towards a dominant, closely related language.

Contextual Hallucination In extractive question answering, models are instructed and constrained to extract answer spans directly from the given context. We define *contextual hallucination* as generating an answer that does not exist as a substring in the source text, distinguishing span selection errors from complete grounding failures (where entire context is ignored). To quantify this, we calculate the *answer presence rate*, the percentage of generated answers found verbatim in the context.

As shown in Figure 4, Gemini-3-Pro outperforms other models by significant margin, indicating its higher instruction following capabilities. While most of the SEA languages remain stable, Burmese (my) and Khmer (km) exhibit a significant *grounding collapse* across the other four models. This degradation creates a sharp divergence along script lines. Latin-based languages usually maintain a high answer presence rate, with models consistently adhering to the source text. In contrast, languages utilising complex abugida scripts (Burmese, Khmer, Lao, and Thai) show a marked reduction in grounding, especially for open-weights models. Because these languages possess native numeral systems, models frequently generate Western Arabic digits even when the context text uses traditional script-specific numerals.

5.4 Tokenizer Analysis

In Figure 5, we compare the relative token fertility across SEA languages using Gemini-3-Pro, Gemma-3 and GPT-5 tokenizers. We observe a shared pattern of high efficiency for Latin-script languages (e.g., Indonesian, Malay, Vietnamese), where all models align closely with the English baseline. However, a distinct performance gap emerges in non-Latin scripts. While all models exhibit increased fertility for languages such as Burmese, Khmer, and Lao, Gemini-3-Pro and Gemma-3 consistently demonstrate superior compression compared to GPT-5. This disparity is most pronounced in Lao, where GPT-5 suffers from extreme fragmentation with a fertility rate approaching 7.0, whereas Gemini-3-Pro has 2.3. These results indicate that while Gemini-3-Pro and Gemma-3 maintain reasonable efficiency across scripts, GPT-5's tokenisation strategy may incur disproportionately high computational costs and latency for non-Latin Southeast Asian languages.

6 Conclusion

We introduced a novel multilingual benchmark, DEEPSEA, designed to evaluate language models across 12 SEA languages, spanning both high- and low-resource tiers. Our extensive evaluation of frontier proprietary models and open-weights models establishes strong baselines on this benchmark and reveals a significant performance disparity among low-resource languages. Further, we identified a critical reliance on English representations, where models consistently perform better when utilising English as an intermediate step rather than operating directly in the target language. We also uncovered distinct failure modes in low-resource settings, such as cross-lingual lexical interference and grounding collapse in abugida scripts. We hope these findings, alongside the release of the DEEPSEA benchmark, will drive the development of next-generation models, which are more inclusive of the SEA region.

Limitations

While DEEPSEA provides a robust benchmark for evaluating SEA languages, it faces two primary limitations. Firstly, our dataset construction primarily relies on translating English-centric (or Indonesian-centric for IndoMMLU) benchmarks into target languages. Although we employed professional translators and a strict quality control process, this approach may inadvertently introduce translation artifacts, potentially favouring models that have been trained heavily on translated data rather than native text. Secondly, despite our efforts to span diverse resource tiers, our coverage is limited to 12 distinct languages. Southeast Asia is a region of immense linguistic complexity, hosting hundreds of dialects and extremely low-resource languages that fall outside the scope of this work.

Acknowledgments

We thank Lucas Bandarkar, Honglin Yu and the multilinguality team in Google Research Australia for insightful discussions on this work. We are grateful to Dinesh Tewari for helping in working with the professional translation vendors. We would also like to thank Grace Chung for her constant guidance and unwavering support.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Alham Fikri Aji and Trevor Cohn. 2025. LORAXBENCH: A multitask, multilingual benchmark suite for 20 Indonesian languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17421–17446, Suzhou, China. Association for Computational Linguistics.

Anthropic. 2026. Claude Sonnet 5 system card. System card, Anthropic.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564, Toronto, Canada. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Abdelrahman Eldesokey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, Aditi Gupta, Afitab Iyigun, Afonso Simplício, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akshay Ramesh, Aleksei Dorkin, and 319 others. 2025. Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures. *Preprint*, arXiv:2510.24081.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark

for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, Haonan Wang, Jiaheng Liu, Yongchi Zhao, Xiachong Feng, Xin Mao, Man Tsung Yeung, Kunat Pipatanakul, Fajri Koto, Min Si Thu, and 22 others. 2025. Sailor2: Sailing in southeast asia with inclusive multilingual llms. *CoRR*, abs/2502.12982.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Gemma Team. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Google DeepMind. 2025a. Gemini 3 Flash model card. Model card, Google DeepMind.

Google DeepMind. 2025b. Gemini 3 Pro model card. Model card, Google DeepMind.

Google DeepMind. 2026. Gemini 3.7 Flash model card. Model card, Google DeepMind.

Dieuwke Hupkes and Nikolay Bogoychev. 2025. Multiloko: a multilingual local knowledge benchmark for llms spanning 31 languages. *CoRR*, abs/2504.10356.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Dahyun Kim, Sukyung Lee, Yungi Kim, Attapol Rutherford, and Chanjun Park. 2025. Representing the under-represented: Cultural and core capability benchmarks for developing Thai large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4114–4129, Abu Dhabi, UAE. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 1–46. Association for Computational Linguistics.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025. SeaExam and SeaBench: Benchmarking LLMs with local multilingual questions in Southeast Asia. In

Findings of the Association for Computational Linguistics: NAACL 2025, pages 6134–6151, Albuquerque, New Mexico. Association for Computational Linguistics.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Chunlan Ma, Ayyoob Imani, Haotian Ye, Renhao Pei, Ehsaneddin Asgari, and Hinrich Schuetze. 2025. Taxi1500: A dataset for multilingual text classification in 1500 languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 414–439, Albuquerque, New Mexico. Association for Computational Linguistics.

Jann Railey Montalan, Jimson Paulo Layacan, David Demitri Africa, Richell Isaiah S. Flores, Michael T. Lopez Ii, Theresa Denise Magsajo, Anjanette Cayabyab, and William Chandra Tjhi. 2025. Batayan: A Filipino NLP benchmark for evaluating large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31239–31273, Vienna, Austria. Association for Computational Linguistics.

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Choon Meng Tan, Brandon Ong, and 11 others. 2025. SEA-LION: southeast asian languages in one network. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP-AACL 2025, Mumbai, India, December 20-24, 2025*, pages 512–526. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2026. GPT-5.6 system card. System card, OpenAI.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Imanol Schlag, Marzieh Fadaee, Sara Hooker, Antoine Bosselut, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, and 40 others. 2025. INCLUDE: evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William-Chandra Tjhi. 2025. SEA-HELM: southeast asian holistic evaluation of language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 12308–12336. Association for Computational Linguistics.

Kiet Van Nguyen, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020. A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024. SeaEval for multilingual foundation models: From cross-

lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

DEEPSEA: A Multilingual Evaluation Benchmark for Southeast Asia
(Appendix)

Dataset	Existing	Newly Created
XQuAD	vi, th	id, fil, ceb, ilo, jv, km, lo, ms, su, my
CrossSum	th, vi, id, my	fil, ceb, ilo, jv, km, lo, ms, su
IndoMMLU	id	th, vi, fil, ceb, ilo, jv, km, lo, ms, su, my
MGSM	th	vi, id, fil, ceb, ilo, jv, km, lo, ms, su, my
WMT24++	vi, th, id, fil	ceb, ilo, jv, km, lo, ms, su, my

Table 4: Overview of the data created and reused from existing collections, showing ISO 639-1 language codes.

Task	Auto-MQM	Human Eval.
MGSM	0.86	90.75%
CrossSum	0.80	95.93%
XQuAD	0.80	95.11%
WMT24++	0.88	-
IndoMMLU	0.95	-
Average	**0.86**	**93.93%**

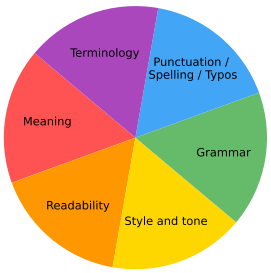

Figure 6: (Left) Translation Quality Assessment. Auto-MQM scores ≥ 0.80 are considered professional quality, and Human Eval percentages represent the rate of accurate translations. (Right) Specific dimensions that are considered in Auto-MQM evaluation.

A DEEPSEA Construction Details

We distinguish the existing publicly available languages from the additional human translations created for DEEPSEA construction in Table 4. Further, the detailed data statistics are provided in Table 5.

B DEEPSEA Quality Check

In order to optimize quality of translations in the benchmark, we implemented a dual-verification process comprising Multidimensional Quality Metrics (MQM) and a meta human review. The Auto-MQM framework evaluates translations across multiple issue dimensions, as illustrated in the pie chart in Figure 6 (right). Complementing this, the meta human review serves as the human validation layer where native experts evaluated translation accuracy by sampling 5% to 10% of unique strings for each task across all languages.

The evaluation results are presented in the table in Figure 6 (left). Under the Auto-MQM framework, translations across all tasks achieved professional quality standards. Due to WMT24++ and IndoMMLU exceeding our automated quality thresholds, we exempted these tasks from the meta review. For the remaining tasks, the human evaluation accuracy is consistent with the automated metrics, which further validates the fidelity of the newly generated data in DEEPSEA benchmark.

C Performance Comparison in Zero-shot and Few-shot Prompting

To assess the effectiveness of in-context learning across different model architectures and language groups, we conducted a comparative analysis of zero-shot and few-shot (three-shot) prompting on mathematical reasoning, scientific reasoning, question answering (QA), and abstractive summarization. The results are illustrated in Figure 7.

Our results reveal a distinct divergence based on task type. On reasoning-heavy tasks (mathematical and scientific reasoning), the few-shot performance is notably worse than zero-shot. We hypothesize that this degradation occurs because the few-shot exemplars provided only direct answers without the intermediate reasoning steps; this likely constrained the models to a direct-answer format, whereas the zero-shot setting better allowed the models to inherently generate the necessary chain-of-thought. Conversely, for the generative tasks (QA and summarization), the few-shot setting proved significantly more effective on most models, likely because the exemplars provided necessary structural guidance and aligned the models with the expected output format helping to prime the model with essential multilingual context to better activate the model's latent cross-lingual knowledge.

Task	Split	Language (code)												
		English (en)	Thai (th)	Vietnamese (vi)	Indonesian (id)	Filipino (fil)	Malay (ms)	Cebuano (ceb)	Ilocano (ilo)	Javanese (jv)	Sundanese (su)	Khmer (km)	Lao (lo)	Burmese (my)
Mathematical Reasoning (MGSM)	test	250	250	250	250	250	250	250	250	250	250	250	250	250
	dev	8	8	8	8	8	8	8	8	8	8	8	8	8
	train	-	-	-	-	-	-	-	-	-	-	-	-	-
Scientific Reasoning (IndoMMLU)	test	735	735	735	735	735	735	735	735	735	735	735	735	735
	dev	15	15	15	15	15	15	15	15	15	15	15	15	15
	train	-	-	-	-	-	-	-	-	-	-	-	-	-
Question Answering (XQuAD)	test	1190	1190	1190	1190	1190	1190	1190	1190	1190	1190	1190	1190	1190
	dev	111	111	111	111	111	111	111	111	111	111	111	111	111
	train	103	103	103	103	103	103	103	103	103	103	103	103	103
Abstractive Summarization (CrossSum)	test	500	101	457	500	500	500	500	500	500	500	500	500	69
	dev	100	100	100	100	100	100	100	100	100	100	100	100	74
	train	100	100	100	100	100	100	100	100	100	100	100	100	100
Machine Translation (WMT24++)	test	-	997	997	997	997	997	997	997	997	997	997	997	997
	dev	-	-	-	-	-	-	-	-	-	-	-	-	-
	train	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 5: Data statistics in DEEPSEA benchmark. The table details the number of instances for each task, broken down by language and data split (test, dev, and train). Note that for the CrossSum task, the splits of th, id, vi, and my languages do not have a one-to-one mapping from the English data.

Model	Mathematical Reasoning			Scientific Reasoning			Question Answering			Abstractive Summarisation			Machine Translation	
	Accuracy ↑			Accuracy ↑			F1 ↑			ChrF ↑			MetricX-24 ↓	
Language	en	high	low	en	high	low	en	high	low	en	high	low	high	low
Gemini-2.5-Pro	97.2	94.5	94.0	84.2	83.8	83.5	83.7	79.8	75.6	35.8	34.7	34.2	2.6	3.7
Gemini-2.5-Flash	97.2	92.9	92.0	83.7	82.4	82.2	82.9	78.2	74.7	33.7	34.0	34.1	2.7	3.8
Gemma-2-27B	93.2	87.9	71.4	67.5	64.3	57.1	85.2	76.3	69.2	35.0	32.7	27.7	3.4	9.3
Gemma-2-9B	91.2	81.9	61.2	65.2	61.1	53.2	79.4	61.7	57.1	33.5	30.4	21.2	3.9	11.6
SEA-LION-v3-9B	94.0	85.0	70.0	71.2	66.0	57.1	83.6	72.7	65.2	33.1	32.4	27.2	3.4	9.3

Table 6: Performance scores of additional models on DEEPSEA, which are divided into English, high- and low-resource non-English language groups.

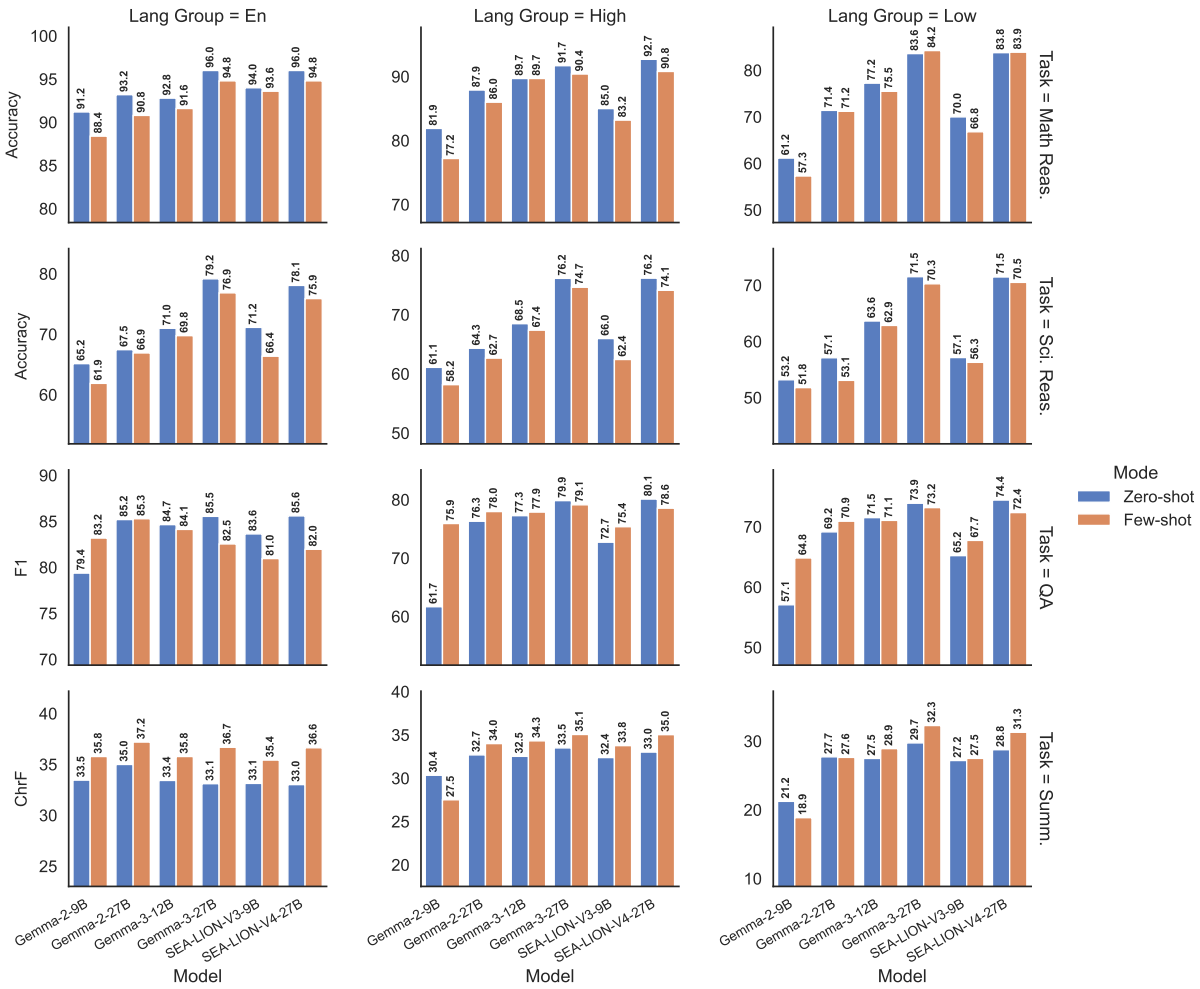

Figure 7: Comparison of various models in zero-shot and few-shot (three-shot) settings.

D Prompts Used In Experiments

The first five prompts below are the default zero-shot prompts used for evaluating the five benchmark tasks. The final three prompts are used specifically for the analysis experiments: the English CoT variants for MGSM and IndoMMLU, and the English-pivot variant for CrossSum. For brevity, we present only the English versions of the reasoning-language prompt templates as representative examples. In the corresponding Native CoT condition, the entire prompt was translated into each target language, and models were explicitly instructed to provide their reasoning in that language.

Prompt D.0.1: Zero-shot Prompt for MGSM

```
You are an expert in solving math reasoning problems. Please solve the following math word problem.
The problem is presented in its original language. Please read it carefully. Please provide
your step-by-step reasoning and solution in the SAME language as the problem. Please conclude
with the final answer in the format: The final answer is <answer>NUMERICAL_ANSWER</answer>

{problem}
```

Prompt D.0.2: Zero-shot Prompt for IndoMMLU

```
You are an expert in solving scientific reasoning problems. Please solve the following multiple
choice problem. The problem is presented in its original language. Please read it carefully.
Please provide your thinking process in the SAME language as the problem. Please conclude with
the final answer in the format: The final answer is <answer>OPTION</answer>

The following are multiple choice questions (with answers) about {subject}.

Question: {question}
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}
```

Prompt D.0.3: Zero-shot Prompt for CrossSum

```
You are an expert multilingual assistant specializing in text summarization. Your task is to
read the following document written in English and generate a concise summary of it in target
language. The summary MUST be only ONE sentence and capture the absolute main idea or most
significant takeaway of the document. The summary must be factually consistent with the source
document.

English Document:
{document}

Please provide the ONE-SENTENCE summary in {target_language} below:

{target_language} Summary:
```

Prompt D.0.4: Zero-shot Prompt for XQuAD

```
You are an expert in solving question answering problems. Please solve the following problem.
Given the context and question, you should answer the question by extracting the answer
*directly* from the context. The context and question are presented in its original language.
Please read it carefully. Please provide your thinking process and answer in the SAME language
as the problem. Please conclude with the final answer in the format: The final answer is
<answer>ANSWER</answer>

Context:
{context}

Question:
{question}
```

Prompt D.0.5: Zero-shot Prompt for WMT24++

```
You are a professional Southeast Asia language translator, tasked with providing translations
suitable for use in Southeast Asia region. Your goal is to accurately convey the meaning and
nuances of the original English text while adhering to target language grammar, vocabulary, and
cultural sensitivities. Produce only the target translation, without any additional explanations
or commentary.

Please translate the following English text into {target_language}:
{input_text}

{target_language} Translation:
```

Prompt D.0.6: Zero-shot Prompt for MGSM with English Reasoning

You are an expert in solving math reasoning problems. Please solve the following math word problem. The problem is presented in its original language. Please read it carefully. Please provide your step-by-step reasoning and solution in **English**. Please conclude with the final answer in the format: The final answer is <answer>NUMERICAL_ANSWER</answer>

{problem}

Prompt D.0.7: Zero-shot Prompt for IndoMMLU with English Reasoning

You are an expert in solving scientific reasoning problems. Please solve the following multiple choice problem. The problem is presented in its original language. Please read it carefully. Please provide your thinking process in **English**. Please conclude with the final answer in the format: The final answer is <answer>OPTION</answer>

The following are multiple choice questions (with answers) about {subject}.

Question: {question}
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}

Prompt D.0.8: Zero-shot Prompt for CrossSum with English Pivot

You are an expert multilingual assistant specializing in text summarization. Your task is to read the following document written in English and generate a concise summary of it. The summary MUST be only ONE sentence and capture the absolute main idea or most significant takeaway of the document. The summary must be factually consistent with the source document. You should first generate the summary in English, then translate it into target language.

You MUST follow this template EXACTLY. Do not add any extra text outside the tags.

<English_Summary>
[A concise, single-sentence summary of the document in English]
</English_Summary>
<Target_Summary>
[The translation of the single-sentence summary into target language]
</Target_Summary>

Target Language: {target_language}

<Source_Document>
{document}
</Source_Document>